Flexible Affix Classification for Stemming Indonesian Language

Reina Setiawan[1,2], Aditya Kurniawan[1,2], Widodo Budiharto[1,2],
Iman Herwidiana Kartowisastro[2], Harjanto Prabowo[2]
[1]Computer Science Department, School of Computer Science
[2]Doctor of Computer Science Department, BINUS Graduate Program
Bina Nusantara University
Jakarta, Indonesia
reina@binus.edu

***Abstract*—The stemming is the process to derive the basic word by removing affix of the word. The stemming is tightly related to basic word or lemma and the sub lemmas. The lemma and sub lemma of Indonesian Language have been grown and absorb from foreign languages or Indonesian traditional languages. Our approach provides the easy way of stemming Indonesian language through flexibility affix classification. Therefore, the affix additional can be applied in easy way. We experiment with 1,704 text documents with 255,182 tokens and the stemmed words is 3,648 words. In this experiment, we compare our approach performance to the confix-stripping approach performance. The result shows that our performance can cover the failure in stemming reduplicated words of confix-stripping approach.**

Keywords—stemming; indonesian language; information retrieval;

I. INTRODUCTION

Stemming is the process to derive a basic word by removing affix of the word. Stemming process is a part of pre-processing text document in information retrieval. The aim of stemming is to increase accuracy of text retrieval [1] The stemming is also needed in compressing text algorithm [2]. In stemming English, Porter stemming algorithm is an algorithm that is simpler than the prior stemming algorithm from Lovin-type stemmers. Porter's algorithm reduces complexity of rule in suffix removal. Hence, this algorithm becomes the standard for English stemmer and provides model for processing of other languages [3].

This paper is focused in stemming Indonesian language (bahasa) that has been growing in lemma. In year 1988, the first edition of Indonesian Language Dictionary - Kamus Besar Bahasa Indonesia (KBBI) released 62,100 lemmas. The second edition in 1991 released 68,000 lemmas, the third edition in 2001 released 78,000 lemmas and the fourth edition in 2008 released 90,000 lemmas [4]. The growth of lemma in Bahasa motivates to create an algorithm of stemming which is more flexible and simple to adapt the enrichment of lemma.

Currently, there are prior several algorithms of stemming Indonesian. Nazief and Adriani algorithm uses a confix-stripping approach with the foundation rule [[[DP+]DP+]DP+] root-word [[+DS][+PP][+P]], Vega algorithm uses iteration to determine and eliminate the affixes from a word. Algorithm from Arifin and Setiono removes up to two prefixes and up to three suffixes. Similar to Vega, this algorithm uses iteration in its process. Asian, William, and Tahaghoghi have improved the confix-stripping approach from Nazief and Adriani. Meanwhile Arifin, Mahendra, Ciptaningtyas also enhanced the confix-stripping approach [5][6][7]. Our approach focuses on affix removal by flexibility affix classification. The affix removal is a type of stemming algorithm [8]. Our approach provides an easy way of stemming Indonesian and more flexible in enhancement. The result of the new approach is compared to result of Nazief and Adriani algorithm. The performance of our approach is better than CS stemmer, especially to stem the reduplicated word.

II. INDONESIAN LANGUAGE AFFIX

Indonesian language (Bahasa) affix consists of prefix, infix, suffix, and confix [9]. The affix is a morpheme that is added to a word to create a new word. The prefix is an affix that is added at the beginning of the word. The infix is an affix that is inserted into the word and the suffix is an affix that is added at the end of the word. Meanwhile, confix is a combination of prefix and suffix in a word. In Bahasa, affix can be added in verb, adjective, adverb or noun. Fig. 1. illustrates the affix concept in Bahasa.

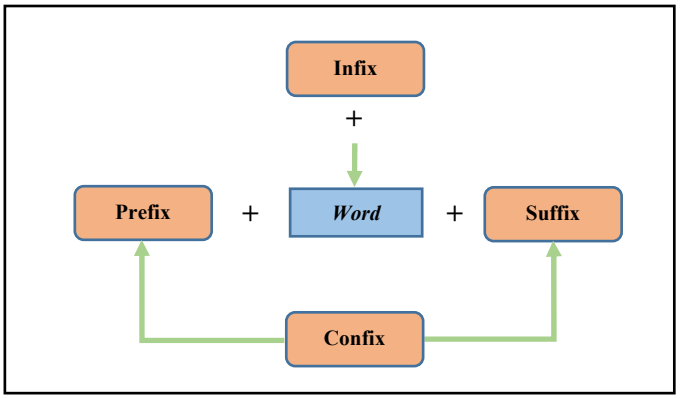

Fig. 1. Affix Concept in Bahasa

978-1-4673-9749-0/16/$31.00 ©2016 IEEE

A. Prefix

Prefix is a morpheme that is added at the beginning of the word to form a new word. There are eight basic types of prefix: 'meng', 'per', 'ber', 'ter', 'di', 'ke', 'peng', and 'se' [10]. The formula of the basic types of prefix is:

$$\{basic\ prefix\} + \{root\ word\} \quad (1)$$

E.g. a word 'mengikat', the formula is 'meng' + 'ikat' and the word is a verb, meaning 'tie'. 'meng' is the prefix and 'ikat' is the root word. Table I shows several examples of the basic types of prefix.

TABLE I. THE EXAMPLES OF THE PREFIX

Basic prefix	Word	Root word	Meaning
'meng'	'mengikat'	'ikat'	tie
'per'	'perpanjang'	'panjang'	renew
'ber'	'berlari'	'lari'	run
'ter'	'terkecil'	'kecil'	smallest
'di'	'dikenang'	'kenang'	remembered
'ke'	'kedua'	'dua'	second
'peng'	'pengikat'	'ikat'	tape
'se'	'sebagai'	'bagai'	as

The morpheme of the several basic prefixes can be changed to become another form, called allomorph and the process is named morphophonemic process. Furthermore, the allomorph is part of type of the prefix and the general formula of prefix is implemented in allomorph. The formula of the allomorph prefix is:

$$\{allomorph\ prefix\} + \{root\ word\} \quad (2)$$

E.g. a word 'melawan', the formula is 'me' + 'lawan' and the word is a verb, meaning 'against'. 'me' is the prefix and 'lawan' is the root word. The prefix 'me' is an allomorph of the prefix 'meng'. However, the prefix 'meng' is not only has one allomorph, 'me'. It has several allomorphs, such as 'men', 'mem', 'meny', and 'menge'. This condition applies to certain basic types of the prefix. Table II shows all of the allomorphs and the rules. The root words that do not comply with the rules in table II, the applicable prefix is the basic prefix.

TABLE II. THE ALLOMORPH AND THE RULES

Basic Prefix	Allomorph Prefix	Rule	Word	Root word	Meaning
'meng'	'me'	root word begins with letter 'l' / 'm' / 'n' / 'r' / 'y' / 'w' / 'ny' / 'ng'	'melawan'	'lawan'	against
'meng'	'men'	root word begins with letter 'd' / 't' / 'c' / 'j' / 'sy'	'mendaki'	'daki'	climb
'meng'	'mem'	root word begins with letter 'b' / 'p' / 'f'	'membeli'	'beli'	buy
'meng'	'meny'	root word begins with letter 's'	'menyadari'	'sadar'	realize
'meng'	'menge'	root word is one syllable	'mengecek'	'cek'	check
'per'	'pe'	root word begins with letter 'r' or the end of the first syllable is 'er'	'peruncing'	'runcing'	sharpening
'per'	'pel'	only or specific root word	'pelajar'	'ajar'	learner
'ber'	'be'	root word begins with letter 'r' or the end of the first syllable is 'er'	'bekerja'	'kerja'	work
'ber'	'bel'	only for specific root word	'belajar'	'ajar'	study
'ter'	'te'	root word begins with letter 'r' or the end of the first syllable is 'er'	'terasa'	'rasa'	feel
'peng'	'pe'	root word begins with letter 'l' / 'm' / 'n' / 'r' / 'y' / 'w' / 'ny' / 'ng' / 'c' / 'j' / 'sy'	'penyanyi'	'nyanyi'	singer
'peng'	'pen'	root word begins with letter 'd' / 't'	'pendobrak'	'dobrak'	burglar
'peng'	'pem'	root word begins with letter 'b' / 'p' / 'f'	'pembeli'	'beli'	buyer
'peng'	'peny'	root word begins with letter 's'	'penyerta'	'serta'	accompanying
'peng'	'penge'	root word is one syllable	'pengecek'	'cek'	checker

Moreover, the process of the prefix additional to the root word can assimilate the first letter of the root word therefore certain prefixes must be replaced with the first original letter to derive the root word. The first original letters are 'k', 't', 's', and 'p'. Table III shows the rule of the first original letter additional.

TABLE III. THE RULE OF ORIGINAL LETTER ADDITIONAL

Rule	Word	Prefix	Root word	Meaning
Prefix 'meng' or 'peng' is replaced with original letter 'k'	'mengenang'	'meng'	'kenang'	remember
Prefix 'men' or 'pen' is replaced with original letter 't'	'menimbang'	'men'	'timbang'	weighing
Prefix 'meny' or 'peny' is replaced	'menyapu'	'meny'	'sapu'	sweeping

with original letter ‘s’				
Prefix ‘mem’ or ‘pem’ is replaced with original letter ‘p’	‘memukul’	‘mem’	‘pukul’	hit

In Bahasa, the prefix can be combined with another prefix to form another word. The formula is:

$$\{1^{st}\ prefix\} + \{2^{nd}\ prefix\} + \{root\ word\} \quad (3)$$

The first prefix consists of ‘meng’ and ‘di’, while the second prefix is ‘per’. This condition also applies to all of the allomorphs. E.g. a word ‘memperpanjang’, the formula is ‘mem’ + ‘per’ + ‘panjang’. ‘mem’ as the first prefix is the allomorph prefix of basic prefix ‘meng’. ‘per’ is the second prefix and ‘panjang’ is the root word. Word ‘memperpanjang’ is a verb, meaning ‘extend’. Table IV shows the combination of the prefixes.

TABLE IV. COMBINATION OF THE PREFIXES

1st Prefix 1	2nd Prefix	Word	Root word	Meaning
‘meng’	‘per’	‘memperpanjang’	‘panjang’	extend
‘di’	‘per’	‘diperkaya’	‘kaya’	enriched

B. *Infix*

Infix is a morpheme that is inserted into the root word. There are four kind of infix: ‘em’, ‘er’, ‘el’, ‘in’. The infix is inserted as the end of the first syllable of the root word. Therefore the formula is:

$$\{1^{st}\ letter\ of\ root\ word\} + \{infix\} + \{remaining\ letters\} \quad (4)$$

E.g. a word ‘telunjuk’. It is a noun and meaning ‘fore finger’. The word ‘telunjuk’ is derived from root word ‘tunjuk’, meaning point. The formula is ‘t’ + ‘el’+ ‘unjuk’. ‘t’ is the first letter of root word ‘tunjuk’, ‘el’ is the infix, and ‘unjuk’ is the remaining letters of the root word. Table V shows the examples of the infixes.

TABLE V. THE EXAMPLES OF THE INFIXES

Root word	Infix	Word	Meaning
‘gembung’	‘el’	‘gelembung’	bubble
‘sabut’	‘er’	‘serabut’	fiber
‘kilau’	‘em’	‘kemilau’	shiny
‘kerja’	‘er’	‘kinerja’	performance

C. *Suffix*

Suffix is a morpheme that is added at the end of the word. There are three kind of suffix: ‘kan’, ‘an’, ‘i’. The formula of the suffix is:

$$\{root\ word\} + \{suffix\} \quad (5)$$

E.g. a word ‘tayangan’. It is a noun and meaning ‘the show’. The word is formed by word ‘tayang’ + ‘an’. ‘tayang’ is the root word and ‘an’ is the suffix. Table VI shows the examples of the suffixes.

TABLE VI. THE EXAMPLES OF THE SUFFIXES

Root word	Suffix	Word	Meaning
‘tawar’	‘kan’	‘tawarkan’	offer
‘tawan’	‘an’	‘tawanan’	hostage

D. *Confix*

Confix is a combination between prefix and suffix to form a new word that is derived from root word. The rules of prefix, the allomorph of prefix, and combination of prefixes also apply in confix concept. The formula of the confix are:

$$\{prefix\} + \{root\ word\} + \{suffix\} \quad (6)$$

$$\{1^{st}\ prefix\} + \{2^{nd}\ prefix\} + \{root\ word\} + \{suffix\} \quad (7)$$

E.g. a word ‘menggunakan’. It is a verb and meaning ‘using’. The formula is ‘meng’ + ‘guna’ + ‘kan’. ‘meng’ is the basic prefix, ‘guna’ is the root word, and ‘kan’ is the suffix. Another example is word ‘membutuhkan’. It is a verb and means ‘need’. The formula is ‘mem’ + ‘butuh’ + ‘kan’. ‘mem’ is the allomorph from prefix ‘meng’. ‘butuh’ is the root word and ‘kan’ is the suffix.

Moreover, there are additional several rules of confix. There is the prefix that is disallowed to combine with certain suffix. This rule is showed in table VII.

TABLE VII. DISALLOWED PREFIX-SUFFIX COMBINATION

Prefix	Disallowed Suffix
‘meng’	‘an’
‘per’	‘an’
‘ber’	‘i’
‘ter’	‘an’
‘di’	‘an’
‘ke’	‘kan’

Another rule is about ordering of combination prefixes, e.g. the combination of prefix ‘meng’ and ‘per’. ‘meng’ is the first prefix and ‘per’ is the second prefix. It is disallowed to place ‘per’ as first prefix and ‘meng’ as the second prefix. This condition also applies to the allomorph prefixes. The ordering rule is showed in TABLE VIII.

TABLE VIII. THE ORDERING RULE OF COMBINATION PREFIXES

1st Prefix	2nd Prefix	Suffix
‘meng’ or ‘di’	‘per’	‘kan’ or ‘i’
‘meng’ or ‘di’ or ‘ter’	‘ber’	‘kan’

E. *Reduplicated*

There is another affix form, named reduplicated. Reduplicated is iteration of word with a new meaning. Reduplicated has several form with the formula are:

$$\{prefix\} + \{root\ word\} + \text{‘-’} + \{root\ word\} \quad (8)$$

$$\{prefix\} + \{root\ word\} + \text{‘-’} + \{root\ word\} + \{suffix\} \quad (9)$$

$$\{root\ word\} + \text{‘-’} + \{root\ word\} + \{suffix\} \quad (10)$$

E.g. a word ‘puji-pujian’, the meaning is ‘laudation’. ‘puji’ is the root word that is iterated with the suffix in the second root word.

F. *Particle and Possessive Pronoun*

There are others suffixes that are derived from particle and possessive pronoun. Particle consists of word ‘kah’, ‘lah’, ‘tah’, and ‘pun’ while possessive pronoun consists of word ‘ku’, ‘mu’, ‘nya’. The particle is used to give affirmation to the root word. The formula is same as suffix formula. E.g. a word ‘pergilah’, meaning ‘please go’, is derived from a root word ‘pergi’ and the particle ‘lah’. Another example is a word ‘mobilku’, meaning ‘my car’ is derived from a root word ‘mobil’ and the possessive pronoun ‘ku’.

III. PROPOSED METHOD

A. *Flexible Affix Classification*

The wide variety of affixes, especially in prefix and suffix encourage our approach to create the classification of the basic prefixes including the allomorph prefixes and suffixes based on number of letter of the prefix and number of letter of the suffix. Therefore, we use the number of letter to name the classification. E.g. the prefix2 classification consists of prefixes with two letters: ‘ke’, ‘di’, ‘se’, ‘me’, ‘pe’, ’be’, ’te’. Prefix ‘ke’, ‘di’, and ‘se’ are the basic prefixes meanwhile ‘me’, ‘pe’, ‘be’, and ‘te’ are the allomorph prefixes. The advantage of our approach is the flexibility of the affix additional. E.g. the prefix such as ‘kese’, ‘sepe’, ‘keber’, ‘keter’, ‘teper’, ‘berse’, ‘seper’, ‘pemer’, ‘pember’, ‘berpen’ are the additional prefix in the fourth Indonesian Language Dictionary - Kamus Besar Bahasa Indonesia (KBBI). This additional is easy to apply in our approach. Since all of the infix consists of two letter, we only have one classification of the infix in current. However, if there is additional of another form of infix (the number of letter is more than two), our approach can implement easily through form of this classification. Table IX and Table X show the classification.

TABLE IX. PREFIX CLASSIFICATION

Classification	Prefix
Prefix2	‘di’, ‘ke’, ‘se’, ‘me’, ‘pe’, ‘be’, ‘te’
Prefix3	‘ber’, ‘bel’, ‘ter’, ‘per’, ‘pel’, ‘pem’, ‘pen’, ‘mem’ ‘men’
Prefix4	‘meng’, ‘meny’, ‘peng’, ‘peny’, ‘kese’, ‘sepe’
Prefix5	‘menge’, ‘penge’, ‘diper’, ‘diber’, ‘keber’, ‘keter’, ‘seper’, ‘berse’, ‘pemer’, ‘teper’
Prefix6	‘memper’, ‘member’, ‘mempel’, ‘pember’, ‘berpen’

TABLE X. SUFFIX CLASSIFICATION

Classification	Suffix
Suffix3	‘kan’, ‘kah’, ‘lah’, ‘tah’, ‘pun’, ‘nya’
Suffix2	‘an’, ‘ku’, ‘mu’
Suffix1	‘i’

B. *Algorithm*

In our approach, we use root word dictionary to validate the word and the word is called token. Our algorithm consists of six steps. First, we check the token to the root word dictionary. If the token is a root word, then we store token as a root word. Second, we check the reduplicated word by checking of the hyphen mark ‘-‘. The first word of reduplicated word can be a root word. If it is not a root word, we process to elimination the prefix. The third step is to check possibility of the confix form. Since confix is combining of prefix and suffix, we check the suffix and the prefix of the word. The next step is to check possibility of prefix, suffix and infix. We eliminate the prefix or suffix or infix depends on the finding.

In every step, we always check the token to the root word dictionary. This process is to confirm result of the elimination of the affix is accomplished. We divide the algorithm in six sub modules: sub module check_dictionary, check_prefix, prefix, check_suffix, suffix, additional, and delete. Sub module check_dictionary validates the token to the root word dictionary. It returns a value that is used as a flag. The sub module check_prefix checks the existence of the prefix in the prefix classification. Sub module prefix eliminates the prefix using sub module delete. It is possible to call sub module additional to add the assimilated letter of the word. Similar with the sub module check prefix, the sub module check_suffix checks the existence of the suffix in the suffix classification and calls sub module delete to eliminate the suffix. Furthermore, the confix processes suffix flow and prefix flow. Fig. 2 illustrates the hierarchy chart of the sub modules.

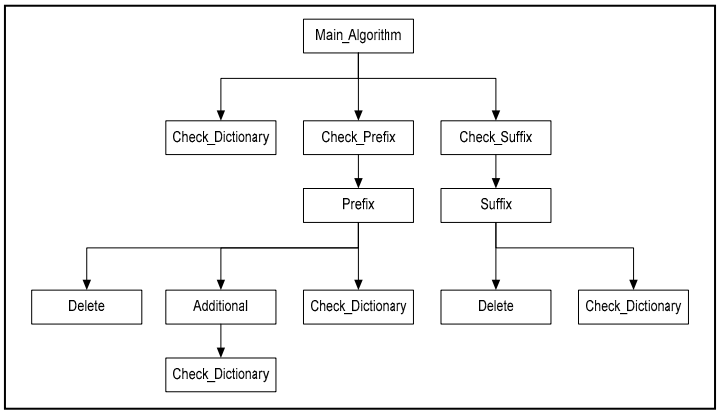

Fig 2. The Hierarchy Chart of The Sub Modules

Our approach applies all of the formulas except disallowed prefix-suffix combination since in our algorithm, we use iteration process to check the affix. Besides the formulas, the approach implement two additional rules of the Indonesian language affix:

Rule 1: If the word is only a letter than it cannot be added by the affix.

E.g. a word ‘memberi’ means ‘give’. Since ‘i’ is only a letter, it cannot be formulated as ‘member’ + ‘i’, although prefix ‘member’ is part of prefix6 classification. It is formulated as ‘mem’ + ‘beri’, therefore ‘mem’ is part of prefix3 classification and ‘beri’ is the root word.

Rule 2: If the word is the reduplicated word than the first word is taken and possible to be processed the prefix

The pseudo code of the algorithm is

```
set and classify the prefixes and suffixes
set the infix
sub module check_dictionary
    check the existence of the token in the root word dictionary
    return flag
end
sub module delete
    remove certain letter(s) from the token
end
sub module additional
    check length of the token
    check the prefix is in assimilated word category
    do check_dictionary with parameter original_letter+token
end
sub module prefix
    do delete
    do check_dictionary
    do additional (if necessary)
end
sub module suffix
    do delete
    do check_dictionary
end
sub module check_prefix
    check prefix classification
    do prefix
end
sub module check_suffix
    check suffix classification
    do suffix
end
main_algorithm
        do check_dictionary
        process reduplicated (if necessary)
        process confix (if necessary)
        process prefix (if necessary)
        process suffix (if necessary)
        process infix (if necessary)
end
```

C. *Experiment*

Furthermore, we code the algorithm by python programming language. We do the pre-processing text document completely in three steps; tokenize the text document, stop-word to remove the common word including words that are not in Indonesian standard, and stemming. The text document, stop-word and the root word dictionary are store in MySQL database. The sample text document is discussion forum from Bina Nusantara University learning management system. There are 1,704 posts in discussion forum with 255,182 tokens that has been processed stop-word removal. The number of stemmed words is 3,648 words. The stemmed words consist of 1,195 words with prefix, 1 word with infix, 696 words with suffix, 1,505 words with confix, and 251 reduplicated words.

This is an example of the discussion forum: "Menanggapi topik no 1 dan no 2 saya berpendapat : 1. Computer Based Information System (CBIS) merupakan sistem pengolah data menjadi sebuah informasi yang berkualitas dan dipergunakan untuk suatu alat bantu pengambilan keputusan. 2. Terdapat 4(empat) komponen utama dari CBIS, yaitu: Hardware : Perangkat yang digunakan untuk menunjang aktifitas sistem. Software : Komponen yang digunakan untuk menjalankan, memerintahkan, memberikan instruksi pada hardware untuk mengolah data. Database : Tempat penyimpanan hasil olahan data. Network : Sistem penghubung antar komputer untuk saling berbagi informasi dan data Terima Kasih,".

We also test the text document with confix-stripping stemmer from Nazief and Adriani through web services from Faculty of Computer Science, University of Indonesia. The url of the web services is http://fws.cs.ui.ac.id/WebServices/. Moreover, the text document is stemmed manually based on the fourth Indonesian Language Dictionary - Kamus Besar Bahasa Indonesia (KBBI).

IV. RESULT AND DISCUSSION

There are three results of the experiment. First is the result of our approach, the second is the result of confix stripping stemmer from Nazief and Adriani through http://fws.cs.ui.ac.id/WebServices/, and the third is the result from manually stemmed based on KBBI. After we do the stop-word and stemming process, the accuracy of performance is showed in table XI.

TABLE XI. THE ACCURACY OF PERFORMANCE

Category	Nazief and Adriani Approach	Our Approach
Prefix	1,132 words	1,131 words
Infix	1 word	1 word
Suffix	683 words	686 words
Confix	1,413 words	1,410 words
Reduplicated	none	251 words

The percentage of the accuracy is described in fig. 3.

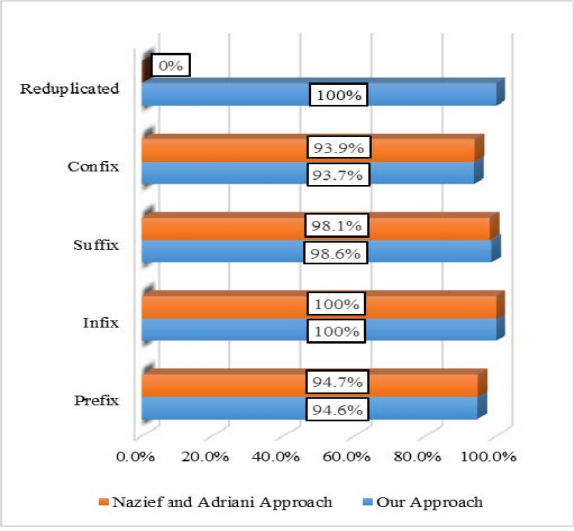

Fig. 3. The Accuracy of Performance

The reduplicated words cannot be stemmed by Nazief and Adriani approach, e.g. 'berbulan-bulan' meaning 'months', 'terburu-buru' meaning 'rush', 'berbeda-beda' meaning 'diverse', 'fakta-fakta' meaning 'facts' and others. Our approach can stem the reduplicated words refer to the *Rule 2*. Example word 'berbulan-bulan', regarding to *rule 2*, we take the first word 'berbulan' and then we eliminate the prefix. 'berbulan' consists of prefix 'ber' and root word 'bulan' meaning 'month'.

In our experiment, we find couple words that are combining two words to become a word with confix. The first word is a word with prefix and the second word is a word with suffix. E.g. word 'ketidakpastian' meaning 'uncertainty', 'ketidakmampuan' meaning 'incompetence', 'ketidakjujuran' meaning 'dishonesty', 'ketidakpercayaan' meaning 'mistrust', 'ketidaksabaran' meaning 'impatience', 'ketidaktahuan' meaning 'unknowing' and others. All of these words cannot be stemmed by Nazief and Adriani approach. In our approach, we can create a new classification easily. We create Prefix7 for the prefix 'ketidak'. The word 'ketidakpastian' consists of prefix 'ketidak', root word 'pasti', and suffix 'an'. Beside capability of reduplicated stem, the flexibility is the advantage of our approach. The comparative of both these approaches is showed in table XII.

TABLE XII. THE COMPARATIVE OF THE APPROACHES

Category	Nazief and Adriani Approach	Our Approach
Reduplicated Stem	No	Yes
Flexibility	No	Yes

However, our approach faces several difficult words, e.g. word 'mengurus'. The word can be stemmed in two ways: 'meng' + 'urus', meaning 'handle' or 'meng' + 'kurus', meaning 'become thin'. Other words are 'mencapai' and 'memakan'. A word 'mencapai' can be stemmed as 'men' + 'capa' + 'i' or 'men' + 'capai' and a word 'memakan' can be stemmed as 'me' + 'makan' or 'mem' + 'akan'. To overcome this problem, the algorithm needs to be improved by considering the meaning of the word through semantic approach.

V. CONCLUSION

The result of stemming from 1,704 text documents in discussion forum with Nazief and Adriani approach compare to our approach shows that the performance of our approach is better than confix-stripping stemmer from Nazief and Adriani. The analysis shows that the advantages of our approach can cover the shortcoming of Nazief approach. The shortcoming is the failure of stemming of reduplicated word. The other advantage is the easy way of adding affixes by flexibility affix classification method in stemming Indonesian language.

REFERENCES

[1] D. Sharma, M. Cse, "Stemming algorithms: a comparative study and their analysis," in International Journal of Applied Information Systems (IJAIS), vol. 4, no. 3, pp. 7-12, 2012.

[2] A. Sinaga, Adiwijaya, H. Nugroho, "Development of word-based text compression algorithm for Indonesian language document," in 3rd International Conference on Information and Communication Technology (ICOICT), 2015, pp. 450-454, DOI: 10.1109/ICoICT.2015.7231466, Bali.

[3] P. Willett, "The Porter stemming algorithm: then and now," in Emerald Insight, vol. 40, iss: 3, 2006, pp. 219-223.

[4] "Kamus Besar Bahasa Indonesia," 4th ed., Departemen Pendidikan Nasional, 2008.

[5] M. Adriani, J. Asian, B. Nazief and et al., "Stemming Indonesian: a confix-stripping approach," in ACM Transactions on Asian Language Information Processing, vol. 6, 2007, pp. 1-33.

[6] J. Asian, H. Williams and S. Tahaghoghi, "Stemming Indonesian," in Conferences in Research and Practice in Information Technology Series, vol. 38, 2005, pp. 307-314.

[7] A.Z. Arifin, I.P.A.K. Mahendra and H.T. Ciptaningtyas, "Enhanced confix stripping stemmer and ants algorithm for classifying news document in Indonesian language," in International Conference on Information & Communication Technology and Systems, 2009, pp. 149-158.

[8] W. B. Frakes, C. J. Fox, "Strength and similarity of affix removal stemming algorithms," ACM SIGIR Forum, vol. 37, 2003, pp. 26-30.

[9] E. Waridah, "EYD ejaan yang disempurnakan & seputar kebahasa-Indonesiaan," Ruang kata, 2015.

[10] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, A. M. Moeliono, "Tata Bahasa Baku Bahasa Indonesia," 3rd ed., Balai Pustaka, 2003.